SPEECH UTTERANCE CLASSIFICATION MODEL TRAINING WITHOUT MANUAL TRANSCRIPTIONS

Ye-Yi Wang[1], John Lee[2] and Alex Acero[1]

[1]Speech Technology Group, Microsoft Research
[2]Spoken Language Systems Group, MIT Computer Science and Artificial Intelligence Laboratory

ABSTRACT

Speech utterance classification has been widely applied to a variety of spoken language understanding tasks, including call routing, dialog systems, and command and control. Most speech utterance classification systems adopt a data-driven statistical learning approach, which requires manually transcribed and annotated training data. In this paper we introduce a novel classification model training approach based on unsupervised language model adaptation. It only requires wave files of the training speech utterances and their corresponding classification destinations for modeling training. No manual transcription of the utterances is necessary. Experimental results show that this approach, which is much cheaper to implement, has achieved classification accuracy at the same level as the model trained with manual transcriptions.

1. INTRODUCTION

Speech utterance classification has recently been widely applied to a variety of spoken language understanding tasks. Call routing is one of the most common applications [1-3]. Other applications include dialog systems and command and control [4]. Most speech utterance classification systems adopt a data-driven statistical learning approach, which requires manual transcriptions of speech utterances and annotations of classification destinations for the utterances. Data transcription and annotation are time-consuming and expensive, and have become a bottleneck for the rapid development of spoken language applications.

Recently many researchers have been investigating different learning algorithms to address the problem. Most of the work focuses on active learning and unsupervised/semi-supervised learning algorithms [5-7] that reduce the requirement of classification destination annotation. However, in many real application scenarios, transcriptions of speech utterances are more difficult to obtain than the classification destinations. One such scenario is the Wizard-of-Oz data collection, in which a wizard interacts with users on behalf of an automated system by choosing a classification destination after hearing the users' utterance. In the case of call-routing, applications are often developed to replace an existing manual system. Data can be collected by recording the users' utterances and the corresponding routing destinations selected by live agents in the manual system. At the initial stage of a system deployment, live agents are often involved in monitoring the system and correcting its classification errors to ensure a smooth transition from using a human operator. In all the scenarios, the classification destinations of the users' utterances can be obtained for free by recording the wizards' or agents' actions. Manual transcription of speech utterances remains as the only major bottleneck in model training and tuning.

In [8, 9], phone sequence based classification systems have been proposed to eliminate the need for manual transcription in classification model training. In both systems, phone n-gram models are iteratively trained and then used by phone recognizers to produce phone sequences. Salient phone subsequences can be selected and used as features by a classifier.

While it is reported that the phone-based classification achieved an accuracy close to that of a word-based model trained on manual transcriptions, it has its own limitations in practical applications: not all speech recognizers support phone recognition; word recognition is often necessary to recover additional attribute values (like phone numbers); sometimes the operational cost is too high when multiple n-gram models and acoustic models are required to cope with the lower detection rate of salient phone sequences or ambiguous salient phone sequences in an utterance [9].

In this paper, we investigate classification model training based on automatic word transcriptions. We first study the effect of using different amounts of seed data for language model adaptation for speech utterance classification. We further study different unsupervised self-adaptation schemes for a word-based language model. We demonstrate the importance of cross-validation in unsupervised adaptation. We illustrate that, even though there is no recognition accuracy improvement over the LM adaptation iterations, classification accuracy may keep improving as long as the errors made by the speech recognizer become more consistent.

The paper is organized as follows: Section 2 gives a brief introduction to the maximum entropy classification algorithm that is used as the statistical classifier in this study. Section 3 describes various language model adaptation algorithms. Section 4 presents experimental settings and results. Section 5 concludes the paper.

2. MAXIMUN ENTROPY CLASSIFIER

Given an acoustic signal A, the task of speech utterance classification is to find the destination class $\hat{C}$ from a fixed set $\mathcal{C}$ that maximizes the conditional probability $\mathrm{P}(C \mid A)$:

$$\begin{aligned}\hat{C} &= \arg\max_{C\in\mathcal{C}} \mathrm{P}(C \mid A) \\ &= \arg\max_{C\in\mathcal{C}} \sum_{W} \mathrm{P}(C \mid W, A)\mathrm{P}(W \mid A) \\ &\approx \arg\max_{C\in\mathcal{C}} \sum_{W} \mathrm{P}(C \mid W)\mathrm{P}(W \mid A) \\ &\approx \arg\max_{C\in\mathcal{C}} \mathrm{P}\left(C \mid \arg\max_{W} \mathrm{P}(W \mid A)\right)\end{aligned} \quad (1)$$

Here W represents a possible word sequence that is uttered in A. The first approximation in Eq. (1) assumes that class C depends only on the word sequence and not on the acoustics A. Many practical systems also make the second Viterbi approximation by adopting a two-pass approach, in which an automatic speech recognition (ASR) engine obtains the best hypothesis of W from A in the first pass; and a classifier takes W as input and identifies its destination class. The text classifier models the conditional distribution $\mathrm{P}(C \mid W)$ and picks the destination according to Eq. (1). There are many different ways to implement a text classifier. We used a Maximum Entropy (ME) classifier [10] for the study in this paper.

A ME classifier builds the conditional distribution on a set of features $\mathcal{F}$. Each feature is a function of C and W. The classifier selects a conditional distribution $\mathrm{P}(C \mid W)$ that maximizes the conditional entropy $H(C \mid W)$ from a family of distributions, with the constraint that the expected count of a feature predicted by the conditional distribution equals to the empirical count of the feature observed in the training data:

$$\sum_{W,C} \tilde{\mathrm{P}}(W) \cdot \mathrm{P}(C \mid W) \cdot f_i(W,C) = \sum_{W,C} \tilde{\mathrm{P}}(W,C) \cdot f_i(W,C), \quad \forall f_i \in \mathcal{F} \tag{2}$$

here $\tilde{\mathrm{P}}$ stands for empirical distributions over the training set.

It has been proven that the maximum entropy distributions that satisfy Eq. (2) have the following exponential (log-linear) form [10]:

$$\mathrm{P}(C \mid W) = \frac{1}{Z_\lambda(W)} \exp\left(\sum_{f_i \in \mathcal{F}} \lambda_i f_i(W,C)\right) \tag{3}$$

where $Z_\lambda(W) = \sum_C \exp(\sum_{f_i \in \mathcal{F}} \lambda_i f_i(W,C))$ is a normalizing constant.

In the application of speech utterance classification, we used binary unigram and bigram features. Formally,

$$f_{u,c}(W,C) = \begin{cases} 1 & \text{when } u \in W \wedge C = c \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where u is a word or a word bigram, c is a destination class.

In Eq. (3), λ_i's are the parameters of the model. They are also known as the weights of feature f_i, which can be optimized from training data. For model training, we applied the stochastic gradient descent algorithm [11], which converges faster than the frequently used generalized iterative scaling (GIS) algorithm [12] in our task.

3. LANGUAGE MODEL ADAPTATION

A straightforward way to train a ME classifier without manual transcriptions is to use ASR transcriptions. Because in-domain transcriptions are not available for language model (LM) training, a general large vocabulary (60K words) dictation trigram language model is used. As shown in the experimental results in the next section, the mismatch of the language model results in an over 50% increase in classification error rate compared to a language model and a classification model trained with manual transcriptions. This is a clear indication that LM adaptation may improve recognition and hence classification accuracy.

One way to adapt the language model involves a small amount of transcribed data. The data is used to train a trigram model, and the resulting model is then interpolated with the large vocabulary trigram model. The interpolation weights are estimated with an automatically transcribed development set. This is a supervised adaptation.

The supervised adaptation approach still requires transcribed data. An alternative is self-adaptation, or unsupervised adaptation [13]. In this case, the wave files of speech utterances are first recognized with the large vocabulary trigram model. The recognized strings are then used to train a domain specific language model. This new language model is interpolated with the original language model used in recognition. Then, iteratively, the newly interpolated model is used to perform recognition again, and the recognition results in turn are used to adapt the original language model.

One problem with the self-adaptation mechanism in [13] is that the recognition errors are fed back to the new language model in the same trigram context. This feedback reinforces the errors in the context and makes it difficult to improve recognition accuracy over iterations. For example, if the word sequence "a b c" is misrecognized as "a b d" in the first iteration, then it is almost hopeless to recover from this error in the subsequent iterations because the error in the first iteration boost the wrong trigram probability in LM adaptation. In this paper we introduce a two-fold cross-validation unsupervised self-adaptation mechanism. The training wave files are randomly partitioned into two disjoint sets, A and B. The original large vocabulary trigram model is used to recognize utterances in A. It is then adapted with the recognized text for the recognition of the utterances in B. The text recognized from B is in turn used to adapt the original language model to recognize again the utterances in A. This process iterates until there is no improvement of classification accuracy on the development set. In the example above, because the word "c" may have different bigram and trigram context in data sets A and B, the cross-validation may alleviate the error reinforcement problem.

4. EXPERIMENTAL RESULTS

We conducted experiments with the ATIS data [14]. The original purpose of ATIS is not for speech utterance classification. However, since the data is broadly accessible to the research community, the results we report with ATIS data can be easily reproduced. We followed the practice in [15] to use the main database table name in the reference SQL query for each utterance as the classification destinations for the utterance. This resulted in 14 different classes. We used the ATIS2 and ATIS3 Category A data (utterances that can be interpreted without context information) for training (5798 utterances), ATIS3 1993 and 1994 Category A test set (914 utterances) for testing, and the ATIS3 development set (410 utterances) for tuning the interpolation parameters and the stopping criterion described in the previous section. In all the experiments we used the recognizer that was

provided as part of the Microsoft Speech API (SAPI) without adaptations to its acoustic model.

To see the potential classification accuracy we can expect, a maximum entropy classification model was trained using all the manually transcribed training data. A trigram language model was also trained using the same data for the recognition of the test utterances. Table 1 shows the best-scenario ASR and classification error rates (CER).

Input to Classifier	Test WER	Test CER
Manual Transcript	0%	4.81%
ASR Output	4.82%	4.92%

Table 1. Performance of the ME classifier on text inputs and speech inputs for the ATIS domain. Both the classifier and the LM for ASR are trained from in-domain manual transcriptions.

In the first experiment, we investigated the effect of LM adaptation with small amounts of in-domain data on the recognition and classification results. Table 2 shows the results for up to 400 randomly selected manually transcribed utterances. The baseline 60K dictation trigram model yielded a 7.44% classification error rate (CER), which is a 51% increase over the 4.92% CER of the best scenario in Table 1.

Adapt. set size	Train WER	Test WER	Test CER
0 (baseline)	18.23%	16.93%	7.44%
100	11.66%	11.34%	5.58%
200	10.02%	10.46%	6.02%
300	9.13%	9.73%	5.35%
400	8.59%	9.60%	5.58%

Table 2. Impact on the WERs and the CERs of using small amounts of manually transcribed data for training the language model. The transcribed utterances were used to adapt the 60K vocabulary dictation trigram LM using linear interpolation.

We applied the unsupervised self-adaptation algorithm [13] to adapt the large vocabulary trigram model with the training wave files. The adapted language model was then used to recognize the training data and the recognized text was used to train the classifier. Table 3 shows the word error rates on the training and test set, as well as the classification error rates on the test set. This self-adaptation mechanism reduces the classification error rate by 26% over the baseline, and it outperforms the approach of LM adaptation with partially transcribed in-domain data in Table 2.

Iteration	Train WER	Test WER	Test CER
1	13.49%	10.18%	5.47%
2	13.38%	10.04%	5.91%
3	**13.44%**	**9.90%**	**5.47%**
4	13.49%	9.93%	5.47%

Table 3. WERs and CERs with unsupervised LM self-adaptation. The row in bold face is the model with the best classification accuracy on development set.

It is interesting to note that the WERs of the training set are about 30% higher than the corresponding WERs of the test set. This indicates that improvement can be achieved if we correctly address the error feedback problem in the self-adaptation mechanism as originally introduced in [13]. This leads to the third experiment, in which we applied the self-adaptation algorithm with two-fold cross-validation. Table 4 shows the word error rates on the training and test set, as well as the classification error rates on the test set.

The training set word error rates in Table 4 are much lower than those in Table 3. In fact, they are in the same range as the word error rates of the test data. This shows that the two-fold cross-validation reduces the error feedback problem in LM self-adaptation. The improvement in the training set recognition accuracy results in training data that better matches the test condition. This accounts for the improvement in the test set classification error rates, which are significantly lower than those in Table 3.

Iteration	Train WER	Test WER	Test CER
1	10.76%	10.06%	5.58%
2	11.03%	10.25%	5.24%
3	11.15%	10.49%	5.03%
4	**11.16%**	**10.36%**	**4.92%**
5	11.17%	10.41%	4.92%
6	11.15%	10.40%	4.81%
7	11.14%	10.41%	5.25%
8	11.17%	10.35%	5.14%

Table 4. WERs and CERs with unsupervised LM self-adaptation and two-fold cross-validation. The row in bold face is the model with the best CER on the development set.

Another important observation in Table 4 is that, although there is no improvement of recognition accuracy after the first iteration (a similar observation was reported in [13]), the test set classification error rate keeps improving until iteration 6. Using the development set CER as the stopping criterion, the model in iteration 4 was chosen. Its CER is the same as the best scenario, when all ATIS manually transcribed training data are used. Our hypothesis is as follows. Once the word error rate is lower than a certain level, further improvement in word error rate may not matter very much. The important issue here is what kind of errors the recognizer makes. As long as the errors are consistent, the classifier can be robustly trained to handle the errors. As an example, about a quarter of the occurrences of the word "flight" was misrecognized as "Floyd," and the maximum entropy classifier learned to include "Floyd" as an important feature for the "FLIGHT" class.

To further test this hypothesis, we looked at the conditional entropy of the distributions of the substitution and deletion errors made by the recognizer, which measures the consistency of recognition errors:

$$H = -\sum_{w} \mathrm{P}(w) \sum_{v} \mathrm{P}(v \mid w) \cdot \log \mathrm{P}(v \mid w) \qquad (5)$$

Here $\mathrm{P}(w)$ is the error distribution over the reference vocabulary, and $\mathrm{P}(v \mid w)$ is the recognizer's confusion distribution for reference word w, which includes $v = \varepsilon$ for deletion errors.

Table 5 lists the conditional entropies through the LM self-adaptation iterations. The drop of entropy correlates strongly with the drop of CERs in Table 4. In iteration 6, when the entropy is at a minimum, the classification error also reaches the lowest level.

This strongly indicates that consistency in ASR errors helps reduce CERs.

Iter.	H	Iter.	H	Iter.	H
1	1.839	4	1.748	7	1.739
2	1.780	5	1.748	8	1.739
3	1.750	6	1.731		

Table 5. Conditional entropy of the ASR substitution and deletion error distribution steadily drops in the initial six iterations of LM self-adaptation. This drop correlates to the classification error drops in Table 4.

We believe that, by making ASR errors more consistent in training and test data, the LM self-adaptation with cross-validation helps the classifier achieve the same accuracy as the model trained with manual transcriptions, as Table 1 and Table 4 illustrate.

5. CONCLUSIONS

In this paper we investigate the problem of training speech utterance classification models without manually transcribed data. We have discussed some important application scenarios for this technology. We have introduced various language model adaptation techniques that customize a domain-independent large vocabulary trigram model for a specific domain, and hence improve the speech recognition and speech utterance classification accuracy. We have shown that unsupervised self-adaptation using all waveform training data, with or without cross-validation, outperforms supervised adaptation using small amounts of transcribed data. The cross-validation mechanism in unsupervised LM self-adaptation successfully mitigates the problem of error-feedback and reinforcement. As a result, the training set recognition error rate is reduced by 30%, and hence the classification error on the test set is significantly reduced to the same level as that obtained from the models trained with manually transcribed training data. We have shown that the high classification accuracy obtained was largely due to the LM self-adaptation, which makes the recognition errors more consistent.

REFERENCES

[1] H.-K. J. Kuo, I. Zitouni, E. Fosler-Lussier, E. Ammicht, and C.-H. Lee, "Discriminative Training for Call Classification and Routing." In the Proceedings of International Conference on Speech and Language Processing, Denver, Colorado, 2002.

[2] M. Gilbert, J. Wilpon, B. Stern, and G. D. Fabbrizio, "Intelligent Virtual Agents for Contact Center Automation." *IEEE Signal Processing Magazine*, vol. 22, 2005.

[3] B. Carpenter and J. Chu-Carroll, "Natural language call routing: a robust, self-organizing approach." In the Proceedings of the International Conference on Speech and Language Processing, Sydney, Australia, 1998.

[4] J. R. Bellegarda, "Semantic Inference: a Data-Driven Solution for NL Interaction." In the Proceedings of International Conference on Speech and Language Processing, Denver, Colorado, USA, 2002.

[5] G. Tur, D. Hakkani-Tür, and R. E. Schapire, "Combining active and semi-supervised learning for spoken language understanding." *Speech Communication*, vol. 45, 2005.

[6] H.-K. J. Kuo and V. Goel, "Active Learning with Minimum Expected Error for Spoken Language Understanding." In the Proceedings of Eurospeech, Lisbon, Portugal, 2005.

[7] R. Sarikaya, H.-K. J. Kuo, V. Goel, and Y. Gao, "Exploiting Unlabeled Data Using Multiple Classifiers for Improved Natural Language Call-Routing." In the Proceedings of Eurospeech, Lisbon, Portugal, 2005.

[8] H. Alshawi, "Effective Utterance Classification with Unsupervised Phonotactic Models." In the Proceedings of HLT-NAACL, Edmonton, Canada, 2003.

[9] Q. Huang and S. Cox, "Automatic Call Routing with Multiple Language Models." In the Proceedings of HLT-NAACL 2004 Workshop on Spoken Language Understanding for Conversational Systems and Higher Level Linguistic Information for Speech Processing, Boston, MA, 2004.

[10] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra, "A Maximum Entropy Approach to Natural Language Processing." *Computational Linguistics*, vol. 22, 1996.

[11] H. J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications*: Springer-Verlag, 1997.

[12] J. H. Darroch and D. Ratcliff, "Generalized iterative scaling for lig-linear models." *Ann. Math. Stat.,*, vol. 43, 1972.

[13] M. Bacchiani and B. Roark, "Unsupervised Language Model Adaptation." In the Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, China, 2003.

[14] P. Price, "Evaluation of Spoken Language System: the ATIS domain." In the Proceedings of DARPA Speech and Natural Language Workshop, Hidden Valley, PA, 1990.

[15] C. Chelba, M. Mahajan, and A. Acero, "Speech Utterance Classification." In the Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, China, 2003.